JABUTI-SQL: Um *Benchmark* em Português para Avaliação de Abordagens *Text-to-SQL* em Cenários Reais

Matheus Botelho[1], Guilherme Fonseca[1], Lucas R. Silva[1], Rodrigo M. Gonçalves[1], Yago Lobato[2], Gustavo A. Ribeiro[1], Eduardo C. de Camara[2], Jaide F. C. Zardin[2], Pedro Calais[1], Emerson A. Simoes[1], Allan Sene[4], Julio C. S. Reis[3], Altigran da Silva[2], Marcos André Gonçalves[1]

[1] Depto. de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)
[2] Instituto de Computação - Universidade Federal do Amazonas (UFAM)
[3] Depto. de Informática - Universidade Federal de Viçosa (UFV)
[4] Dadosfera Brasil

{matheusbotelho, guilhermefonseca, emerson.simoes}@dcc.ufmg.br
{lucasrsilvak, rodrigopfmg, gustavo9661, pedrolgcalais}@ufmg.br
{yagobrlobato, eduardo.camara, jaide.zardin, alti}@icomp.ufam.edu.br
allan@dadosfera.ai, jreis@ufv.br, mgoncalv@dcc.ufmg.br

Abstract. *This paper presents JABUTI-SQL (**J**oint **A**nnotated **B**enchmark for **U**ser-centric **T**ext-to-SQL **I**n Real Health Data), a Brazilian benchmark for Text-to-SQL built from real public healthcare data provided by DATASUS, the Brazilian public health data platform. The benchmark integrates multiple heterogeneous sources related to pharmaceutical assistance, public procurement, hospital infrastructure, and regional healthcare organization into a unified relational schema. Unlike synthetic benchmarks, JABUTI-SQL preserves characteristics commonly found in real governmental databases, including incomplete records, semantic inconsistencies, heterogeneous structures, and complex inter-table relations. The benchmark contains 69 Portuguese question–SQL pairs covering three user personas and multiple difficulty levels, aiming to support the evaluation of Text-to-SQL systems under more realistic conditions.*

1. Introdução

O crescimento da disponibilidade de dados públicos e os avanços recentes em modelos de linguagem impulsionaram o interesse por aplicações de *Text-to-SQL*, que consistem em mecanismos capazes de traduzir consultas em linguagem natural para comandos SQL [Katsogiannis-Meimarakis and Koutrika 2023]. Essas abordagens representam um caminho promissor para democratizar o acesso à informação, permitindo que usuários com diferentes níveis de conhecimento técnico explorem bases estruturadas de forma mais acessível [Hong et al. 2024, Pedroso et al. 2025].

No contexto da saúde pública brasileira, o DATASUS[1] disponibiliza um conjunto extenso de informações relacionadas à infraestrutura hospitalar, assistência farmacêutica e compras públicas do Sistema Único de Saúde (SUS). Essas bases, conjuntamente, possuem elevado potencial analítico para aplicações envolvendo, dentre outras possibilidades, monitoramento epidemiológico, gestão pública, auditoria e transparência

[1] https://dadosabertos.saude.gov.br/

governamental [Finkelstein and Borges Junior 2020, Moraes et al. 2025]. Entretanto, o uso efetivo desses dados ainda enfrenta desafios importantes, incluindo heterogeneidade estrutural, inconsistências semânticas, ausência de padronização, dados incompletos e fragmentação entre diferentes sistemas [Silva et al. 2025].

Essas características tornam particularmente desafiadora a construção de *benchmarks* realistas para avaliação de sistemas *Text-to-SQL*, especialmente em língua portuguesa [Pedroso et al. 2025, Fróes and Braghetto 2025]. Embora *benchmarks* amplamente utilizados neste contexto, como Spider [Yu et al. 2018] e BIRD [Li et al. 2023], tenham impulsionado avanços na área, eles frequentemente assumem esquemas mais controlados e bases relativamente limpas quando comparados a cenários governamentais reais.

Com o objetivo de aproximar a avaliação de sistemas *Text-to-SQL* das condições encontradas em aplicações reais, este trabalho apresenta o JABUTI-SQL (***J**oint **A**nnotated **B**enchmark for **U**ser-centric **T**ext-to-SQL **I**n Real Health Data*), um *benchmark* construído a partir de múltiplas bases públicas do DATASUS e disponível publicamente no seguinte link: `https://doi.org/10.5281/zenodo.21048827`. O *benchmark* integra dados relacionados à infraestrutura hospitalar, rede farmacêutica e compras públicas em um esquema relacional unificado, preservando características típicas de ambientes reais, como dados incompletos, ambiguidades semânticas e relações complexas entre entidades.

Além da integração das bases, o trabalho disponibiliza um conjunto de pares pergunta/SQL em português organizados por diferentes níveis de dificuldade e perfis de usuários. Para isso, foram realizados processos de coleta, normalização, integração e documentação dos dados, bem como a construção de consultas representativas de cenários analíticos reais. O recorte temporal adotado compreende o ano de 2025, buscando equilibrar atualidade dos dados e cobertura temporal mínima para análises comparativas. Esperamos o que o JABUTI-SQL possa ser útil para fomentar pesquisas neste contexto.

2. Trabalhos Relacionados

Esta seção apresenta estudos relacionados considerando duas dimensões principais: i) trabalhos que exploram dados públicos do contexto da saúde, enfatizando desafios neste contexto, e; ii) pesquisas relacionadas a tarefa de *Text-to-SQL*.

2.1. Dados Públicos de Saúde

A disponibilização de dados públicos de saúde por meio de plataformas governamentais tem impulsionado uma crescente produção científica voltada à compreensão e à melhoria do sistema de saúde brasileiro [Finkelstein and Borges Junior 2020, Gamarski et al. 2025]. Bases como o Sistema de Informações Hospitalares (SIH), o Cadastro Nacional de Estabelecimentos de Saúde (CNES) e a Base Nacional de Dados de Ações e Serviços da Assistência Farmacêutica (BNAFAR) têm sido exploradas em diferentes estudos, que evidenciam tanto o seu valor analítico quanto os obstáculos inerentes ao seu uso. [Finkelstein and Borges Junior 2020], por exemplo, analisa padrões de utilização e cobertura do SUS a partir de dados hospitalares, ilustrando o potencial dessas fontes para subsidiar avaliações de desempenho do sistema público. Em outra perspectiva, [Silva et al. 2025] investiga os preços praticados por municípios brasileiros na aquisição de medicamentos da atenção primária a partir de dados da BNAFAR, demonstrando que municípios socioeconomicamente mais vulneráveis, especialmente nas

regiões Norte e Nordeste, pagam sistematicamente mais caro por esses insumos. Além de revelar desigualdades estruturais relevantes para a formulação de políticas públicas, o estudo evidencia limitações importantes dos dados originais: registros autodeclarados, ausência de obrigatoriedade de adesão universal e preenchimento incompleto por parte dos municípios, características que impõem barreiras ao uso direto dessas bases e reforçam a necessidade de processos sistemáticos de curadoria e normalização.

Esforços voltados à aplicação de tecnologias de Inteligência Artificial sobre esses dados também começam a surgir na literatura. [Moraes et al. 2025] propõem um agente conversacional para interação com dados de hospitalização do DATASUS, demonstrando a viabilidade de modelos de linguagem de grande porte nesse domínio. Os autores, no entanto, relatam limitações relevantes em consultas analíticas que envolvem múltiplas tabelas, reforçando que a complexidade estrutural e a heterogeneidade semântica das bases governamentais representam desafios ainda em aberto e que justificam iniciativas de curadoria e integração como a proposta neste trabalho.

2.2. Recursos para *Text-to-SQL*

Entre as aplicações viabilizadas por bases de dados bem estruturadas, o *Text-to-SQL*, tarefa de traduzir consultas em linguagem natural para comandos SQL, tem recebido atenção crescente da comunidade de Processamento de Linguagem Natural. Benchmarks como Spider [Yu et al. 2018], WikiSQL e BIRD [Li et al. 2023] desempenharam papel fundamental no avanço da área, ao fornecer grandes coleções de pares pergunta/SQL para treinamento e avaliação de modelos[Fonseca et al. 2026], contribuindo para o desenvolvimento de métodos capazes de lidar com consultas complexas e múltiplos domínios. Entretanto, a maior parte desses recursos foi construída em língua inglesa e sobre esquemas relativamente controlados, com bases limpas, bem estruturadas e de baixa heterogeneidade semântica quando comparadas a cenários reais de aplicação.

No contexto multilíngue, iniciativas recentes adaptaram benchmarks existentes para o português [José and Cozman 2023, Pedroso et al. 2025, Dou et al. 2023], ampliando a cobertura linguística da tarefa. Ainda assim, esses recursos herdam limitações estruturais dos benchmarks originais, especialmente a ausência de bases reais, a falta de integração entre múltiplas fontes e a supressão de ruídos característicos de ambientes de produção. De forma geral, os recursos existentes, tanto para análise de dados públicos de saúde quanto para avaliação de sistemas *Text-to-SQL*, ainda não endereçam de maneira integrada os desafios impostos por bases governamentais reais: heterogeneidades estruturais, ambiguidades semânticas, dados incompletos e relações complexas entre entidades de múltiplas fontes. O presente trabalho busca preencher essa lacuna ao propor uma base de dados curada, integrada e documentada, construída a partir de fontes reais do DATASUS e projetada para suportar múltiplas tarefas analíticas.

3. Construção do *Benchmark* JABUTI-SQL

Esta seção descreve o processo de construção do *Benchmark* JABUTI-SQL[2] (***J**oint **A**nnotated **B**enchmark for **U**ser-centric **T**ext-to-SQL **I**n Real Health Data*), incluindo a coleta, integração e padronização das bases públicas utilizadas na composição do mesmo, conforme ilustrado na Figura 1. O objetivo foi construir um ambiente relacional que

[2]Disponível em: `https://github.com/BotelhoMDS/JABUTI-SQL`.

preservasse características típicas de cenários reais, como heterogeneidade estrutural, dados incompletos e dependências entre múltiplas fontes governamentais.

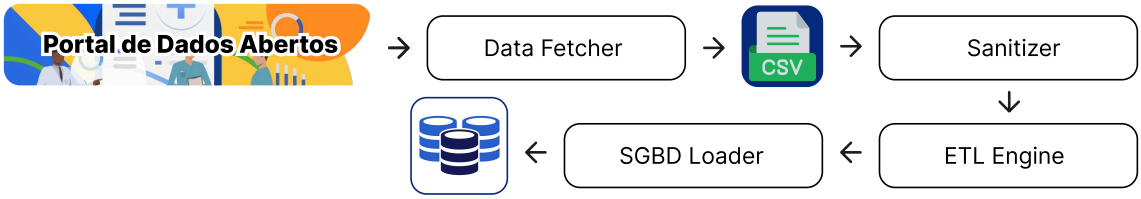

Figura 1. Fluxo de coleta e integração dos dados.

3.1. Fontes de Dados e Coleta

Diferentemente de *benchmarks* sintéticos ou excessivamente controlados, o JABUTI-SQL foi construído a partir da integração de múltiplas bases públicas do DATASUS, preservando características típicas de ambientes governamentais reais, como dados incompletos, ambiguidades semânticas, heterogeneidade estrutural e relações interdependentes entre múltiplas fontes. Para garantir consistência temporal, foi realizado um recorte utilizando os dados referentes ao ano de 2025.

Em suma, o *benchmark* combina três fontes principais. 1) O **Banco de Preços em Saúde (BPS)** reúne informações sobre fornecedores, fabricantes, produtos e gastos públicos, permitindo análises comparativas entre instituições e padrões de aquisição em diferentes níveis geográficos; 2) O conjunto **Hospitais e Leitos** adiciona informações institucionais, geográficas e temporais sobre a infraestrutura hospitalar, contemplando consultas sobre disponibilidade, distribuição e caracterização de leitos, e; 3) A base **BNAFAR**, por sua vez, abrange a disponibilidade de medicamentos, controle de estoque e distribuição regional de insumos farmacêuticos, integrando dados sobre produtos, lotes, validade e quantidades disponíveis. Por fim, para enriquecimento das informações e ampliação da conectividade semântica e geográfica do esquema relacional, foram incorporados dados auxiliares do **Cadastro Nacional de Estabelecimentos de Saúde (CNES)** e das divisões de Macrorregião e Região de Saúde. Essas fontes complementares permitem integrar instituições presentes em diferentes bases e habilitam consultas envolvendo aspectos administrativos, geográficos, demográficos e populacionais, de forma que ao todo o *benckmark* é composto pela junção dessas cinco bases de dados.

As bases foram selecionadas buscando representar diferentes níveis de complexidade estrutural, qualidade dos dados e relevância analítica para tarefas de *Text-to-SQL*. Enquanto o BPS apresenta consultas relativamente simples, a BNAFAR envolve relações mais complexas entre estoque, instituições e logística farmacêutica. A escolha temática também favorece a estratificação por personas: os domínios cobertos permitem formular perguntas acessíveis ao público geral – como gastos públicos, disponibilidade de medicamentos e distribuição regional de recursos – bem como consultas de maior profundidade técnica, voltadas a analistas com conhecimento do esquema e das características internas dos dados.

A coleta foi realizada automaticamente por meio de rotinas de extração desenvolvidas para identificar arquivos disponibilizados em formatos CSV e ZIP no portal de dados abertos. Após a obtenção dos arquivos, os dados foram decodificados e convertidos para um formato relacional unificado, respeitando diferenças de codificação, separadores

Tabela 1. Valores nulos por tabela no repositório de dados.

Tabela	Total de Nulos	% Nulos
instituicao_estoca_produto	699.651.457	44,48%
instituicao	5.833.070	28,96%
endereco	528.861	13,27%
mantenedora_compra_produto	33.078	9,01%
v_endereco_completo	527.222	5,42%
catmat	545	1,07%
v_produto_completo	545	0,44%
demais tabelas	0	0,00%

e padrões estruturais presentes nas bases originais. Como contribuição adicional do estudo e para fins de reprodutibilidade, todos os coletores implementados no estudo estão disponíveis no repositório do projeto[2].

3.2. Integração e Padronização dos Dados

Após a etapa de coleta, os dados passaram pelo processo de limpeza, harmonização e padronização voltado à construção do *benchmark*. Essa etapa incluiu normalização de valores numéricos, padronização de separadores decimais, expansão de siglas e adequação semântica de atributos originalmente pouco descritivos. Também foram realizadas inferências pontuais para preenchimento de informações ausentes quando tais valores podiam ser obtidos de outras relações do esquema.

Apesar do processo de limpeza, parte dos registros permaneceu com valores nulos, especialmente nas tabelas `instituicao_estoca_produto`, `instituicao` e `endereco`, conforme apresentado na Tabela 1. Esses dados ausentes foram deliberadamente preservados, pois refletem características reais dos dados públicos de saúde, frequentemente marcados por inconsistências administrativas, incompletude e heterogeneidade estrutural. Além disso, a manutenção desses registros aproxima o *benchmark* ainda mais de cenários reais de uso e introduz desafios adicionais para sistemas *Text-to-SQL*, que precisam lidar corretamente com consultas envolvendo dados incompletos sem comprometer a qualidade das respostas retornadas.

Em seguida, os dados semi-estruturados são normalizados para tabelas que representam as entidades consideradas chaves para a organização do repositório pelo módulo de *ETL Engine*. Este processo, além de reduzir consideravelmente o consumo de memória do repositório, é de suma importância para o processo de *Text-to-SQL*, pois otimiza as consultas ao reduzir a ambiguidade e a potencial alucinação de modelos de linguagem, permitindo que eles busquem pela informação em uma tabela ou conjunto específico em vez de explorar todo o conjunto de dados [Pourreza and Rafiei 2023].

Por fim, os dados são inseridos em um banco de dados pelo *SGBD Loader*. A Tabela 2 apresenta a volumetria final e a estrutura das entidades geradas, cujos volumes variam de 5 registros, no caso de Região do Brasil, a mais de 143 milhões, em Instituição Estoca Produto, e cujas estruturas oscilam entre 2 e 33 colunas, variações que implicam diferentes graus de complexidade para as consultas realizadas sobre o banco. Durante o processo de integração, o esquema relacional foi reorganizado para reduzir ambiguidades e ampliar a compatibilidade entre as fontes. Em particular, os produtos foram estruturados segundo o código CATMAT, permitindo relacionar itens comprados no BPS a produtos estocados no BNAFAR mesmo diante de informações incompletas ou inconsistentes

Tabela 2. Descrição da estrutura e volume de dados do JABUTI-SQL.

Tabela (Entidade)	Colunas	Qtd. de Dados (Linhas)
Instituição	33	610.321
Mantenedora	4	7.390
Endereço	9	442.672
Município	5	5.570
Região de Saúde	3	439
Macrorregião de Saúde	3	121
Unidade Federativa	4	27
Região do Brasil	2	5
Leitos	17	86.147
Fornecedor	3	1.083
Fabricante	3	751
Produto	5	20.875
CATMAT	3	16.953
Mantenedora Compra Produto	14	26.215
Instituição Estoca Produto	11	143.009.579

entre os conjuntos. O esquema relacional resultante é apresentado na Figura 2.

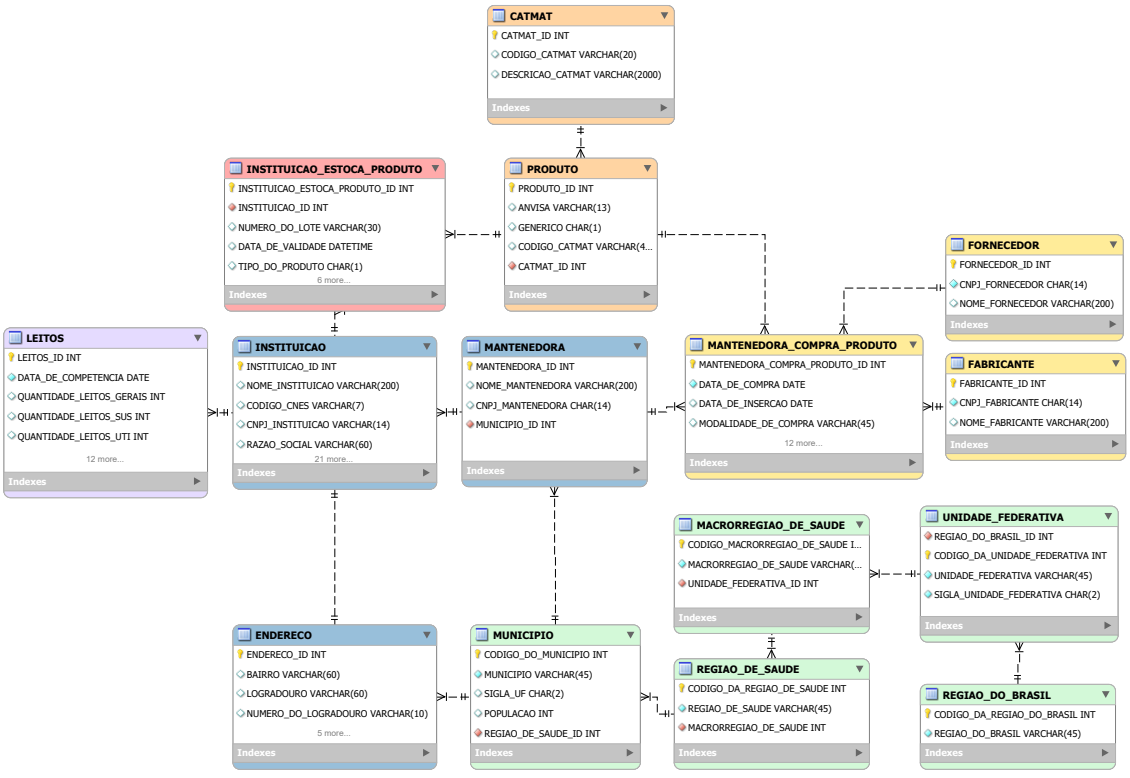

Figura 2. Esquema relacional do JABUTI-SQL.

4. Caracterização do *Benchmark* JABUTI-SQL

A construção de *benchmarks* para tarefas *Text-to-SQL* requer não apenas diversidade estrutural de consultas, mas também representatividade dos cenários reais nos quais os modelos serão aplicados. Nesse contexto, esta seção caracteriza o JABUTI-SQL, detalhando o processo de geração dos pares pergunta/SQL, a definição das personas

utilizadas, a distribuição das consultas por níveis de dificuldade e os principais domínios temáticos contemplados. Também são apresentadas características regionais da base integrada, evidenciando limitações e desafios decorrentes da heterogeneidade e incompletude dos dados utilizados.

4.1. Pares de Perguntas/Consultas

A construção dos pares pergunta/SQL seguiu um *pipeline* semi-automático em duas etapas, inspirado em trabalhos recentes [Camara et al. 2026]. Na primeira, um modelo de linguagem, durante esta etapa do estudo foi explorado o *GPT-5.0*, recebeu o esquema relacional integrado e foi instruído a gerar consultas SQL em três níveis de dificuldade: fácil, médio e difícil. Este processo foi executado a partir com base na taxonomia proposta por [Yu et al. 2018]. Na segunda etapa, as consultas geradas foram fornecidas ao mesmo modelo, que produziu perguntas em linguagem natural correspondentes a cada SQL, respeitando o perfil linguístico de três personas de usuário. Todos os pares resultantes foram validados manualmente para assegurar a correção das consultas e a adequação das perguntas ao perfil de cada persona. As três personas adotadas refletem diferentes graus de familiaridade com o domínio e com o esquema relacional:

- **Usuário leigo:** sem conhecimento de domínio ou de SQL, formula perguntas que refletem necessidades informacionais mais diretas;
- **Médico gestor hospitalar:** com conhecimento do domínio da saúde, porém sem domínio técnico do esquema relacional;
- **Analista de dados do setor de saúde:** especialista com conhecimento do domínio, do esquema relacional e de técnicas de SQL.

O conjunto resultante totaliza 69 pares pergunta/SQL, com 23 perguntas atribuídas a cada persona. Em relação à dificuldade, a distribuição compreende 30 consultas simples ($\approx$ 44%), 30 médias ($\approx$ 43%) e 18 difíceis ($\approx$ 26%). Exemplos de consultas e seus respectivos graus de dificuldade são apresentados na Tabela 3, enquanto a Tabela 4 detalha os componentes SQL presentes nas consultas.

O número de pares é modesto em comparação a benchmarks de larga escala como Spider [Yu et al. 2018] e BIRD [Li et al. 2023], conforme ilustrado na Tabela 6, porém superior ao volume apresentado por [Moraes et al. 2025], que reúne 51 pares em cenário igualmente voltado a dados reais de saúde brasileira. O JABUTI-SQL não busca competir em volume, mas em realismo e representatividade. Cada par foi construído sobre um esquema multi-fonte com dados reais, incompletos e semanticamente heterogêneos, condições que amplificam a dificuldade intrínseca de cada consulta de forma não comparável a ambientes controlados. A estratificação por três personas e três níveis de dificuldade garante cobertura de diferentes perfis de usuário e complexidades SQL, maximizando a diversidade analítica dentro de um escopo cuidadosamente curado. Essa complexidade pode ser observada na Tabela 5, em que os modelos avaliados apresentam quedas expressivas de acurácia ao serem aplicados sobre o JABUTI-SQL em comparação ao benchmark de [Moraes et al. 2025].

4.2. Cobertura Temática

Além da estratificação por dificuldade e persona, as perguntas foram agrupadas por categorias temáticas para caracterizar o escopo analítico do *benchmark*. Como há

Tabela 3. Exemplos de consultas por nível de dificuldade.

Dificuldade	Pergunta	SQL
Fácil	Por tipo de gestão, quantas instituições de saúde estão ativas? Liste primeiro os tipos com mais unidades.	SELECT tipo_de_gestao, COUNT(*) AS total_instituicoes FROM instituicao WHERE motivo_da_desabilitacao IS NULL GROUP BY tipo_de_gestao ORDER BY total_instituicoes DESC
Médio	Qual é a quantidade total de medicamentos disponíveis em estoque por estado?	SELECT ve.sigla_unidade_federativa, SUM(mv.quantidade_do_item_em_estoque) AS quantidade_total_medicamentos FROM mv_estoque_mais_recente mv JOIN instituicao i ON i.instituicao_id = mv.instituicao_id JOIN v_endereco_completo ve ON ve.endereco_id = i.endereco_id WHERE mv.quantidade_do_item_em_estoque $>$ 0 GROUP BY ve.sigla_unidade_federativa ORDER BY quantidade_total_medicamentos DESC, ve.sigla_unidade_federativa
Difícil	Considerando o último registro de leitos de cada instituição, qual é a média de leitos gerais por habitante nos municípios com mais de 500.000 habitantes?	SELECT AVG(mv.quantidade_leitos_gerais ::numeric / mu.populacao) AS media_leitos_por_habitante FROM mv_leitos_mais_recente mv JOIN instituicao inst ON inst.instituicao_id = mv.instituicao_id JOIN endereco e ON e.endereco_id = inst.endereco_id JOIN municipio mu ON mu.codigo_do_municipio = e.municipio_id WHERE mv.quantidade_leitos_gerais IS NOT NULL AND mu.populacao IS NOT NULL AND mu.populacao $>$ 500000

Tabela 4. Distribuição de construções SQL nas consultas do *benchmark*.

Característica	Nº de Consultas	%
`ORDER BY`	60	87,0%
`GROUP BY`	57	82,6%
`JOIN`	39	56,5%
`WITH` (CTEs)	21	30,4%
`HAVING`	18	26,1%
Subconsultas clássicas	6	8,7%

Tabela 5. Resultados de execução exata (EX) por modelo no JABUTI-SQL.

Modelo	EX
Qwen2.5-Coder-32B	9.2%
Qwen2.5-Coder-14B	8.9%
Qwen2.5-Coder-7B	7.7%

Tabela 6. Comparação entre *benchmarks Text-to-SQL*.

Característica	Spider	BIRD	JABUTI-SQL
Idioma	Inglês	Inglês	Português
Dados reais	Não	Parcial	Sim
Dados incompletos	Não	Não	Sim (até 44%)
Fontes integradas	1	1	5
Personas de usuário	Não	Não	Sim (3 perfis)
Pares pergunta/SQL	8.659	12.751	69
Nº med. Joins/SQL	0.5	1.0	1.4
Nº med. Tabelas/Consulta	1.51	2.00	2.30

sobreposição entre categorias, uma mesma pergunta pode estar associada a múltiplos temas, de modo que o total de ocorrências na Tabela 7 excede o número de pares do *benchmark*.

As categorias refletem os principais domínios analíticos suportados pelo esquema. Consultas **geográficas** e de **leitos** permitem avaliar a distribuição regional de recursos hospitalares, frequentemente combinadas entre si. As temáticas **financeira** e de **estoque** têm destaque para aplicações de gestão e auditoria, possibilitando identificar padrões de consumo, anomalias de preço e indícios de má gestão. Consultas sobre **produtos** e **instituições de saúde** complementam essas análises, oferecendo granularidade sobre os itens adquiridos e o perfil dos estabelecimentos. Por fim, a categoria de **caracterização dos dados** suporta consultas voltadas à própria qualidade dos registros, relevante dado o caráter manual e heterogêneo do preenchimento nas bases governamentais.

Tabela 7. Distribuição das perguntas por categoria temática.

Categoria	Ocorrências
Geográfica	18
Leitos	18
Estoque[3]	15
Financeira[4]	12
Produtos	6
Instituições de saúde	3
Caracterização dos dados	3
Total	75

4.3. Dados por região

A Tabela 8 apresenta a distribuição dos registros por região do país. O elevado percentual de registros sem região associada (47,66%) evidencia uma limitação estrutural relevante: a vinculação geográfica depende do preenchimento consistente do código CNES pelas

[1]Ex.: controle de validade, quantidade em estoque.

[2]Ex.: gasto por mantenedoras, lucro por fornecedores.

Tabela 8. Distribuição dos registros por região do Brasil.

Região	Registros	Percentual
Sem Região	68.154.216	47,66%
Nordeste	32.832.712	22,96%
Sudeste	15.046.192	10,52%
Sul	11.294.129	7,90%
Centro-Oeste	8.751.323	6,12%
Norte	6.931.007	4,85%
Total	143.009.579	100,00%

instituições, tornando parte significativa dos registros desconectada das hierarquias regionais utilizadas em consultas analíticas. Ainda assim, os dados cobrem todas as cinco macrorregiões brasileiras, com destaque para o Nordeste (22,96%). Essa heterogeneidade, combinada à presença expressiva de dados incompletos, introduz desafios adicionais para sistemas *Text-to-SQL* em consultas agregadas e filtros geográficos.

5. Considerações Finais

Este trabalho apresentou o JABUTI-SQL, um *benchmark* para avaliação de sistemas *Text-to-SQL* em português do Brasil construído a partir da integração de múltiplas bases públicas reais do DATASUS. O objetivo do *benchmark* é aproximar a avaliação de sistemas *Text-to-SQL* de cenários reais encontrados em bases governamentais, caracterizados por heterogeneidade estrutural, integração entre múltiplas fontes e alto grau de incompletude dos dados.

De forma geral, o JABUTI-SQL consolida informações relacionadas à assistência farmacêutica, compras públicas, infraestrutura hospitalar e organização regional do sistema de saúde brasileiro em um esquema relacional unificado. Isso permite a formulação de consultas analíticas envolvendo múltiplas entidades, relações geográficas e diferentes níveis de complexidade. Diferentemente de *benchmarks* sintéticos ou excessivamente simplificados, o JABUTI-SQL preserva características típicas de ambientes reais, incluindo dados incompletos, inconsistências semânticas, ambiguidades estruturais e dependências complexas entre tabelas.

Além do **potencial de uso** para avaliação de modelos *Text-to-SQL*, o JABUTI-SQL também constitui um recurso relevante para pesquisas envolvendo integração de dados públicos, qualidade de dados, recuperação de informação e aplicações de inteligência artificial em saúde. A natureza heterogênea do esquema permite investigar desafios relacionados à interoperabilidade semântica, tratamento de dados incompletos e consultas analíticas sobre bases governamentais reais. Neste contexto, é importante mencionar que JABUTI-SQL já vem sendo utilizado em pesquisas conduzidas pelo grupo, incluindo estudos sobre geração automática de consultas, avaliação do desempenho de abordagens *Text-to-SQL* em cenários reais. O *benchmark* apresenta limitações no volume de pares pergunta/SQL e na abrangência de domínios cobertos, características inerentes à sua natureza exploratória e ao escopo definido para este trabalho. Como trabalhos futuros, pretende-se ampliar a cobertura temporal e temática do repositório, incorporar novos domínios do DATASUS e expandir o conjunto de pares pergunta/SQL com consultas mais complexas e maior diversidade linguística. Finalmente, espera-se que o JABUTI-SQL contribua para o desenvolvimento de sistemas *Text-to-SQL* mais robustos e melhor adaptados aos desafios encontrados em bases públicas reais do contexto brasileiro.

Uso de Inteligência Artificial

Este trabalho utilizou os modelos Claude, ChatGPT e Gemini para revisão gramatical, tradução do resumo para o inglês, geração de pares de perguntas e consultas e apoio no tratamento de dados, com supervisão e revisão dos autores.

Agradecimentos

INCT-TILD-IAR (#408490/2024-1), FAPEMIG, CAPES, Dadosfera, Embrapii, SEBRAE.

Referências

Camara, E., Zardin, J., Lobato, Y., Fonseca, G., Botelho, M., Simoes, E. A., Silva, L. R., Calais, P., Gonçalves, R., Ribeiro, G., Sene, A., Reis, J. C. S., Gonçalves, M. A., and da Silva, A. (2026). Pairs: Um pipeline para geração automática e curadoria de benchmarks text-to-sql. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD) – Demonstrations*.

Dou, L., Gao, Y., Pan, M., Wang, D., Che, W., Zhan, D., and Lou, J.-G. (2023). Multispider: towards benchmarking multilingual text-to-sql semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12745–12753.

Finkelstein, B. J. and Borges Junior, L. H. (2020). A capacidade de leitos hospitalares no Brasil, as internações no SUS, a migração demográfica e os custos dos procedimentos. *Jornal Brasileiro de Economia da Saúde*, 12(3):273–280.

Fonseca, G., Reis, J. C. S., Botelho, M., Calais, P., Simoes, E. A., Silva, L. R., Camara, E., Zardin, J., Gonçalves, R., Ribeiro, G., Lobato, Y., Sene, A., da Silva, A., and Gonçalves, M. A. (2026). Avaliação sistemática de text-to-sql em português: Um benchmark unificado multi-domínio. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*.

Fróes, K. d. C. and Braghetto, K. R. (2025). Exploring temporal text-to-sql challenges in brazilian portuguese: Lessons from educational data. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 963–969.

Gamarski, R., Castro, F. F. d., Nascimento, J. A. S. d., and Abreu, M. M. d. (2025). A frequência de doenças cardiovasculares na artrite reumatoide no brasil: Estudo de coorte de 10 anos com bancos de dados do DATASUS. *Arquivos Brasileiros de Cardiologia*, 122(2):e20240313.

Hong, Z., Yuan, Z., et al. (2024). Next-generation database interfaces: A survey of llm-based text-to-sql. *arXiv preprint arXiv:2406.08426*.

José, M. A. and Cozman, F. G. (2023). A multilingual translator to sql with database schema pruning to improve self-attention. In *International Journal of Information Technology*, pages 3015–3023.

Katsogiannis-Meimarakis, G. and Koutrika, G. (2023). A survey on deep learning approaches for text-to-sql. *The VLDB Journal*, 32(4):905–936.

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Geng, R., Huo, N., et al. (2023). Can llm already serve as a database interface? a big bench for large-scale

database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36:42330–42357.

Moraes, M., Figueiredo, I., Marques, V., Santos, J., and Manssour, I. H. (2025). From questions to answers: A natural language interface for datasus hospitalization data. In *Escola Regional de Aprendizado de Máquina e Inteligência Artificial da Região Sul (ERAMIA-RS)*, pages 296–299.

Pedroso, B. C., Pereira, M. R., and Pereira, D. A. (2025). Performance evaluation of llms in the text-to-sql task in portuguese. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 260–269.

Pourreza, M. and Rafiei, D. (2023). Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *Advances in neural information processing systems*, 36:36339–36348.

Silva, W. R. O. d., Cavalcanti, I. T. d. N., Peters, J. R., Ferreira, L. E. M. d. S., Santana, R. S., and Leite, S. N. (2025). Preços pagos pelos medicamentos para atenção primária pelos municípios brasileiros. *Revista de Saúde Pública*, 59:1–15.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 3911–3921.